Automatic Interesting Object Extraction From Images Using Complementary Saliency Maps

Haonan Yu[1], Jia Li[2,3], Yonghong Tian[1], Tiejun Huang[1]
[1]National Engineering Laboratory for Video Technology (NELVT), School of EE & CS, Peking University
[2]Key Lab of Intell. Info. Process, Inst. of Comput. Tech., Chinese Academy of Sciences, China
[3]Graduate University of Chinese Academy of Sciences, China
{hnyu,yhtian,tjhuang}@pku.edu.cn,jli@jdl.ac.cn

ABSTRACT

Automatic interesting object extraction is widely used in many image applications. Among various extraction approaches, saliency-based ones usually have a better performance since they well accord with human visual perception. However, nearly all existing saliency-based approaches suffer the integrity problem, namely, the extracted result is either a small part of the object (referred to as *sketch-like*) or a large region that contains some redundant part of the background (referred to as *envelope-like*). In this paper, we propose a novel object extraction approach by integrating two kinds of "complementary" saliency maps (i.e., sketch-like and envelope-like maps). In our approach, the extraction process is decomposed into two sub-processes, one used to extract a high-precision result based on the sketch-like map, and the other used to extract a high-recall result based on the envelope-like map. Then a classification step is used to extract an exact object based on the two results. By transferring the complex extraction task to an easier classification problem, our approach can effectively break down the integrity problem. Experimental results show that the proposed approach outperforms six state-of-art saliency-based methods remarkably in automatic object extraction, and is even comparable to some interactive approaches.

Categories and Subject Descriptors

I.4.6 [**Image Processing and Computer Vision**]: Segmentation – *Pixel classification.*

General Terms

Algorithms, Experimentation

Keywords

Automatic object extraction, Complementary saliency maps, Pixel classification

1. INTRODUCTION

In recent years the number of digital images has grown dramatically. In these images, the truly meaningful parts may be just a small proportion. The nontrivial contents, usually in the form of interesting objects, are sufficient to represent the semantic meanings in most cases and consequently play an important role in many image applications such as content-based retrieval.

Therefore, many methods have been proposed to automatically extract interesting objects. For example, graph-theoretic approaches make use of energy function optimization to solve the extraction problem (e.g., [1, 2]); edge-linking methods, such as [3], connect a subset of the fragments produced by edge detection to form a closed contour for the interesting object, etc. Although these approaches work well in some cases, the tendency to solve the extraction problem with little consideration of human visual perception makes them have undesirable performance under some complicated conditions such as in cluttered images.

Because visual saliency well accords with human visual perception and can be used as one sort of selection mechanisms of the important content, saliency-based approach is proposed recently as an alternative for object extraction. For example, Itti *et al.* [4] combined multiscale features into a single topographical saliency map and adopted a dynamical neural network to select the attended areas that roughly contained the interesting objects. Ma and Zhang [5] generated a contrast-based saliency map and extracted objects by fuzzy growing. Achanta *et al*. [6] outputted a frequency-tuned saliency map and binarized it with an adaptive threshold. Hou and Zhang [7] constructed the saliency map by analyzing the log-spectrum of the image and used a simple threshold to detect pro-objects.

Although these approaches work well to simulate human visual perception, their results usually lack integrity and exactness. That is, the result is either a small part of the object or a large region that contains some redundant part of the background. According to the definition of visual saliency, a region with a higher contrast to its surrounding will be more likely to stand out in the saliency map. This gives rise to dark center areas and over-highlighted edges on a large object (referred to as *sketch-like*), or leads to the redundant detection of local sudden changes in background as a highlighted part (referred to as *envelope-like*).

To solve this problem, we propose a novel interesting object extraction approach using two saliency maps. The two maps, in a complementary manner, are a sketch-like and an envelope-like saliency maps. We simply decompose the extraction process into two sub-processes. The results of the two sub-processes are also somewhat complementary in the sense of exactness, with a high precision and a high recall respectively. We then use the two results as prior knowledge and adopt a simple method for pixel classification. By transferring the complex object extraction task

to an easier classification problem, our approach can effectively break down the integrity problem. Extensive experiments show that the approach outperforms six state-of-art saliency-based ones (i.e., [4, 5, 6, 7, 8, 9]) in automatic interesting object extraction. Moreover, the visual effect of our results is even comparable to some interactive techniques.

The remainder of this paper is organized as follows. Section 2 describes our approach for interesting object extraction. Experimental results are presented in Section 3 and the paper is concluded in Section 4.

2. EXTRACTION APPROACH USING COMPLEMENTARY SALIENCY MAPS

We first give two definitions which will be used later. Suppose the pixel set of an interesting object is denoted as $\mathbf{O}$, then $\mathbf{E}$ is called the *envelope* of the object if $\mathbf{O} \subseteq \mathbf{E}$, and $\mathbf{S}$ is called the *skeleton* of the object if $\mathbf{S} \subseteq \mathbf{O}$. However, in our study, we allow $\mathbf{O} - \mathbf{E} \neq \varnothing$ and $\mathbf{S} - \mathbf{O} \neq \varnothing$ as long as the following two conditions are satisfied: 1) The envelope covers the interesting objects as much as possible while leaving just few redundant background areas. 2) The skeleton contains the most representative parts of the object while including little background.

In the following sub-sections, we first generate two complementary saliency maps and then obtain two complementary results, one for the envelope, and the other for the skeleton. Then we adopt a classification step to extract the exact object. The framework of our approach is shown in Figure 1.

2.1 Estimate the envelope of the object

Firstly, an envelope-like saliency map $\mathbf{M}_{\mathbf{env}}$ is calculated to highlight a rough area as the envelope. Here we construct $\mathbf{M}_{\mathbf{env}}$ by simply weighting two feature maps. The first is *frequency-tuned saliency map* (**FSM**) proposed by Achanta *et al*. [6], and the second is *center-surround contrast map* (**CCM**) derived from Liu *et al*. [10]. The reason of choosing these two feature maps is that **FSM** can output desirable results with very efficient computation while **CCM** can well represent the regional contrast feature and is insensitive to local sudden changes. Besides, as we will show later, **CCM** fails near the image boundary, but **FSM** can work well for the border.

Here we briefly recall the method of **FSM**. Firstly a DoG filter is used for band pass filtering, and then based on the Gaussian blurred image, for any pixel x, the saliency value is computed as:

$$f_{FSM}(x) = \| \mathbf{P}_{aver} - \mathbf{P}(x) \| , \tag{1}$$

where $\mathbf{P}(x)$ is the visual feature at pixel x and $\mathbf{P}_{aver}$ stands for the average of all the features. Each feature vector is chosen as the pixel value in Lab color space. After the computation, we normalize the feature map to [0, 1].

The construction of **CCM** relies on the prior that an interesting object usually distinguishes from its surrounding context. When calculating the center-surround contrast feature, Liu *et al*. empirically set several rectangular templates to match the object region and to represent the strip that surrounds the object. Here, we improve the parameter setting method based on sample data of

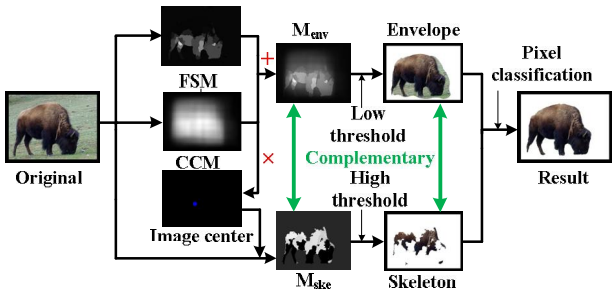

Figure 1. The framework of our approach. Note that the green lines represent complement in the sense of exactness.

object size. According to the data, we choose the most representative parameters to build our rectangular templates. After this, let the template be T, and its equal-size surrounding strip be T_s. One way to measure the difference between T and T_s is to calculate the distance between their color histograms ($\mathbf{T}$, $\mathbf{T_s}$, respectively). Here we use a measurement similar to χ^2 distance:

$$\chi'^2(T, T_s) = \sum_i \left(\frac{\mathbf{T}(i) - \mathbf{T_s}(i)}{\mathbf{T}(i) + \mathbf{T_s}(i)}\right)^2 , \tag{2}$$

where $\mathbf{T}^*(i)$ stands for the ith bin of histogram $\mathbf{T}^*(\mathbf{T}^* \in \{ \mathbf{T}, \mathbf{T_s} \})$. For every pixel x in the image, we obtain several template-strip pairs along with their distances (except for the pixels near the image border). We then pick the max distance and write the relevant T as $\mathcal{T}(x)$:

$$\mathcal{T}(x) = \arg\max_{T(x)} \chi'^2(T(x), T_s(x)) , \tag{3}$$

with its paired T_s as $\mathcal{T}_s(x)$. Then the contrast feature of pixel x is:

$$f_{CCM}(x) \propto \sum_{\{x' | x \in \mathcal{T}(x')\}} d_{xx'} \chi'^2(\mathcal{T}(x'), \mathcal{T}_s(x')) , \tag{4}$$

where the weight $d_{xx'} = e^{-\frac{\|x - x'\|^2}{2\sigma_{x'}^2}}$ is a Gaussian fall off weight with variance $\sigma_{x'}^2$. Finally, we also normalize this feature map to [0, 1].

Given the two feature maps, we construct $\mathbf{M}_{\mathbf{env}}$ to extract the envelope of the object. The saliency value at each pixel x in $\mathbf{M}_{\mathbf{env}}$ is simply calculated as a weighted sum:

$$f_{env}(x) = \alpha_{FSM} f_{FSM}(x) + \alpha_{CCM} f_{CCM}(x) , \tag{5}$$

where $\alpha_{FSM} + \alpha_{CCM} = 1$, $\alpha_{FSM}, \alpha_{CCM} \in [0, 1]$. We set the weights according to the importance of the feature maps. To binarize $\mathbf{M}_{\mathbf{env}}$, we introduce an adaptive threshold $\overline{T}$ which is determined as:

$$\overline{T} = \frac{\alpha_t}{h \times w} \sum_x f_{env}(x) , \tag{6}$$

where h and w are the height and width of the image respectively and α_t is set to a low value to obtain a high recall. Then all the pixels belonging to the envelope can be written as a set $\mathbf{E} = \{ x \mid f_{env}(x) \geqslant \overline{T} \}$. The second row of Figure 2 shows some examples of the envelope.

2.2 Extract the skeleton of the object

To obtain the skeleton, a sketch-like saliency map $\mathbf{M}_{\mathbf{ske}}$ is needed to detect precisely the important parts of the interesting

object. Here we generate $\mathbf{M_{ske}}$ using color spatial-distribution feature [10].

To extract the color spatial-distribution feature, we first calculate the *image center* of an interesting object. Since the center of the object is usually among the most salient parts and thus stands out almost in whatever saliency map, we multiply **FSM** and **CCM** to approximately locate the *image center*. For pixel x, we derive its image-center value as:

$$f_{IC}(x) = (f_{FSM}(x))^{\beta_{FSM}} (f_{CCM}(x))^{\beta_{CCM}}, \quad (7)$$

where β_{FSM} and β_{CCM} are positive weights and are set according to the expected influences of the two saliency maps. Finally, *image center* (h^*, w^*) is defined as:

$$h^* = \sum_x \frac{f_{IC}(x(h_x, w_x))h_x}{\sum_x f_{IC}(x(h_x, w_x))}, \quad (8)$$

with w^* defined similarly. For those images that contain more than one interesting object, this method still works and in this condition *image center* refers to the point locating in the middle of the interesting objects.

Based on the *image center*, $\mathbf{M_{ske}}$ can now be calculated. Depending on the prior [10] that the wider a color distributes, the less possible it is on the interesting object, we can find a representative color by measuring its spatial distribution. These representative colors, usually occupying only parts of the object, can be treated as the skeleton. We cluster the image colors by n Gaussian Mixture Models (GMMs). Suppose $p(i \mid x)$ is the probability that pixel x belongs to model i, and $V(i)$ is the ith model's spatial variance, then the value for pixel x is written as:

$$f_{ske}(x) \propto \sum_i p(i \mid x)(1 - V(i))(1 - D(i)), \quad (9)$$

$$D(i) = \sum_x p(i \mid x) d_x, \quad (10)$$

where d_x is the distance from pixel x to the *image center* and both $D(i)$ and $V(i)$ have been normalized to [0, 1]. Different from [10], our image-center distance can better assign a larger weight to the pixel which is more likely to be salient.

Based on $\mathbf{M_{ske}}$, the skeleton is defined adaptively similar to (6):

$$\mathbf{S} = \{x \mid f_{ske}(x) \geq \frac{\sum_{\{x' \mid x' \in \mathbf{E}\}} f_{ske}(x')}{\mathcal{A}(\mathbf{E})} \varepsilon_t\}, \quad (11)$$

where $\mathcal{A}(\mathbf{E})$ is the area of the envelope and ε_t is set to a high value to gain a high precision. Some skeleton examples are shown at the third row in Figure 2.

2.3 Derive the exact object by classification

The envelope and skeleton are then, used as prior knowledge to finally extract the exact object. Generally speaking, pixels that are not included in the envelope (Fig 2. Second row) have an extremely high probability of belonging to the background. Hence, we label them as background seeds. Meanwhile, pixels which are included in the skeleton (Fig 2. Third row) are highly likely to be parts of the object and we label them as foreground seeds. Then according to the color similarity with the two seed parts, we classify the rest of pixels in the image using a classification method. Here we adopt *color signatures* similar to [11]:

1. **Seeds clustering:** To establish a background KD-tree and a foreground KD-tree for the background seeds and the foreground seeds, respectively. Every node in the trees is a cluster of pixel colors.
2. **Pixel assignment:** For every remaining pixel, to find the nearest tree node in the color space (e.g., by Euclidean Distance) and then assign the pixel to this cluster.
3. **Post-processing:** To connect isolated components or smooth to optimize the result.

After these three steps, an exact object can be extracted[1].

3. EXPERIMENTS

3.1 Dataset and the evaluation metrics

In the experiments, we adopt an image set containing 1000 images which is proposed by [6]. All the 1000 images have exact binary object masks.

We use three frequently used evaluation metrics: *Precision*, *Recall* and *F-measure*. F-measure is defined as:

$$F_\alpha = \frac{(1+\alpha) Precision \times Recall}{\alpha \times Precision + Recall}, \quad (12)$$

where $\alpha = 0.5$.

3.2 Results

In this section, we evaluate the feasibility of our interesting object extraction approach. We first test our envelope and skeleton extraction methods. We compare their results with the object masks, resulting in precision = 0.61, recall = 0.96 for envelope and precision = 0.90, recall = 0.76 for skeleton. This demonstrates that the envelope can well cover the object while the skeleton can well represent the main part of the object.

We then demonstrate the visual effect of our final results. Figure 2 shows some examples. It can be seen that in most cases our approach performs very well. The integrity and exactness of extracted objects is perfect, even though some interesting objects have similar colors with the background or have unclear boundaries (e.g., "Dandelion" and "Hunting duck"). Moreover, despite that "Bicycle" has a complex structure and not uniform color distribution, the whole object is cut out successfully with only a little background. For images that contain multiple interesting objects (e.g., the first and third column), the results are still desirable. Nevertheless, there are also few failures, such as "Boat". It indentifies the reflection in the water as a part of the interesting object, probably due to the inaccurate detection of the skeleton. Finally, we also compare the visual effect with some state-of-art interactive approaches for interesting object extraction. As an example, the comparison with GrabCut [12] is shown in Figure 3. It can be seen that our results are comparable or even better.

To quantitatively evaluate the overall performance, we compare our result with six state-of-art saliency-based approaches: Itti98 [4], Ma03 [5], Achanta09 [6], Hou07 [7], Harel07 [8] and Achanta08 [9]. To optimize some original approaches [4, 5, 8],

[1] For those readers who are interested in this classification method, we refer them to the original papers for the algorithm details.

Figure 2. First row: The original images. Second row: the envelope of the object. Third row: the skeleton of the object. Last row: the final results. It can be seen that our results are robust and edge well-defined, even in some tough cases.

we uniformly binarize their output saliency maps with the method proposed in [6]. The approaches that are optimized either adopted complicated but undesirable binarization methods or outputted only salient points. It can be seen from Figure 4 that our method remarkably outperforms all the other six methods, with a precision of **0.88** and a recall of **0.89**.

Generally speaking, the success of our method is mainly due to that we reconstruct the extraction process by integrating complementary saliency maps and then classifying the pixels based on the two complementary results. We model the extraction problem in a different way and successfully transfer it to an easier classification problem. However, like most other approaches, our approach fails when the contrast between objects and the background is not so obvious, especially when the scenes are complex.

4. CONCLUSION

In this paper, we propose a novel automatic approach to extract interesting objects. Our main contribution is that we successfully break down the integrity problem in most saliency-based approaches. To achieve this goal, we integrate two complementary saliency maps and then exploit the complementary results to classify pixels. From the experimental results, our method outperforms several state-of-art saliency-based methods and is even comparable to some interactive methods. In the future work, we will extend our approach to complex scenes, and we are planning to extract interesting objects from videos.

5. ACKNOWLEDGMENTS

The authors would like to thank Yexiang Xue for the valuable work in the experiments part. This work is supported by grants from the Chinese National Natural Science Foundation under contract No. 60973055 and No. 90820003, National Basic Research Program of China under contract No.2009CB320906, and Fok Ying Dong Education Foundation under contract No. 122008.

6. REFERENCES

[1] C. Li, C. Xu, C. Gui and M. D.Fox, Level Set Evolution Without Re-initialization: A New Variational Formulation. *IEEE CVPR*, 2005.

[2] N. Xu, R. Bansal and N. Ahuja, Object Segmentation Using Graph Cuts Based Active Contours. *IEEE CVPR*, 2003.

[3] S. Wang, T. Kubota, J. M. Siskind, and J. Wang, Salient Closed Boundary Extraction with Ratio Contour. *IEEE PAMI*, Vol. 27, No.4, April 2005.

[4] L. Itti, C. Koch and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE PAMI*, 1998.

[5] Y. Ma and H. Zhang. Contrast-based image attention analysis by using fuzzy growing. *ACM Trans. on Multimedia.*, 2003.

[6] R. Achanta, S. Hemami, F. Estrada and S. Susstrunk. Frequency-tuned Salient Region Detection. *IEEE CVPR*, 2009.

[7] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. *IEEE Conference on Computer Vision and Pattern*, 2007.

[8] J. Harel, C. Koch and P. Perona. Graph-based visual saliency. *Advances in Neural Information Processing Systems*, 2007.

[9] R.Achanta, F.Estrada, P. Wils and S. Susstrunk. Salient region detection and segmentation. *International Conference on Computer Vision Systems*, 2008.

[10] Liu, T, Yuan, Z, N. Zheng, X. Tang and H. Shum. Learning to detect a salient object. *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[11] G. Friedland, K. Jantz and R. Rojas. SIOX: Simple Interactive Object Extraction in Still Images, *Proceedings of the IEEE International Symposium on Multimedia*, 2005.

[12] C. Rother, V. Kolmogorov and A. Blake. "GrabCut" -- Interactive Foreground Extraction using Iterated Graph Cuts, *ACM SIGGRAPH*, 2004.

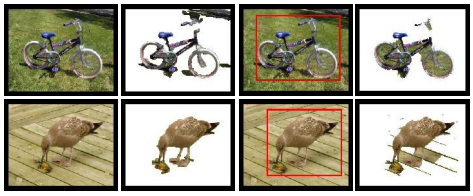

Figure 3. Visual comparison with GrabCut. The second and fourth columns are our and Grabcut's results, respectively.

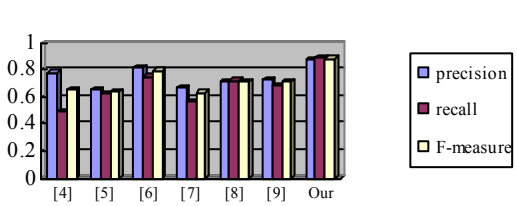

Figure 4. The comparison results with six saliency-based methods. The improvements (%) of F-measure from left to right are: 35.4, 37.5, 10.0, 39.7, 23.9 and 22.2.